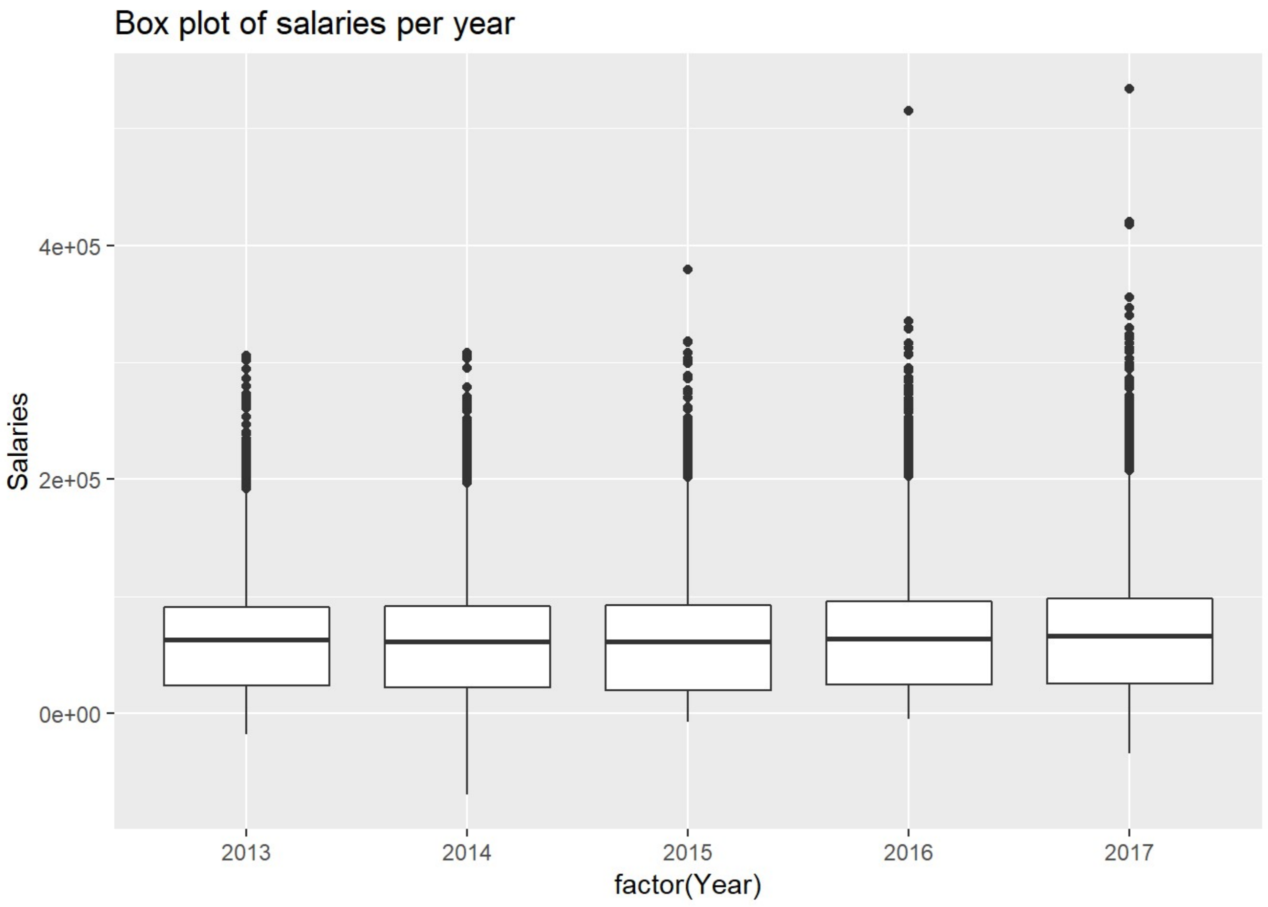
R Notebook

1a) Box-plot for salaries by year.

The boxplot is useful to study the distribution of data and identify outliers. The 25th, 50th and 75th percentiles can be seen in the boxplot. The 50th percentile is the center line in the box that represents the median of data. The points seen in the plot represent the outliers.

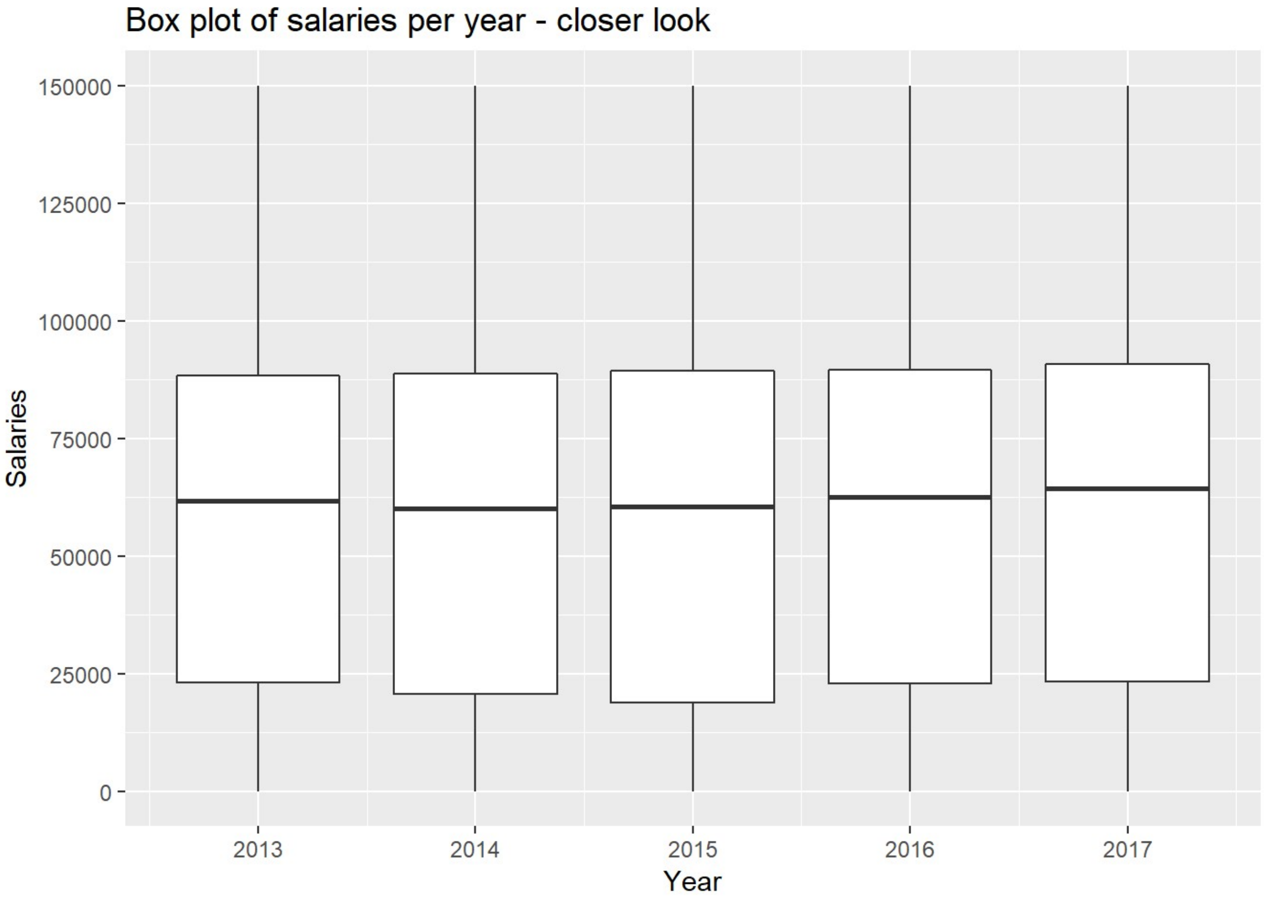
The range of salaries for each year is different. 2017 has the highest salary range as compared with other years. However, the median of salaries is approximately the same for all years. The outliers can be seen distinctly for each year. We shall get more detailed view of the box plot by changing the ylim in the next plot.

```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot()+geom_boxplot(mapping = aes(x=factor(Year),y=Salaries), data=Emp_Comp)+ggtitle
("Box plot of salaries per year")
```



Setting limit for salaries on the y axis to get a better view of the quartiles and median for each year. We can see that the 3rd quartile varies slightly for each year and so is the median.Taking a closer look, the largest non-outlier also varies slightly for each year.

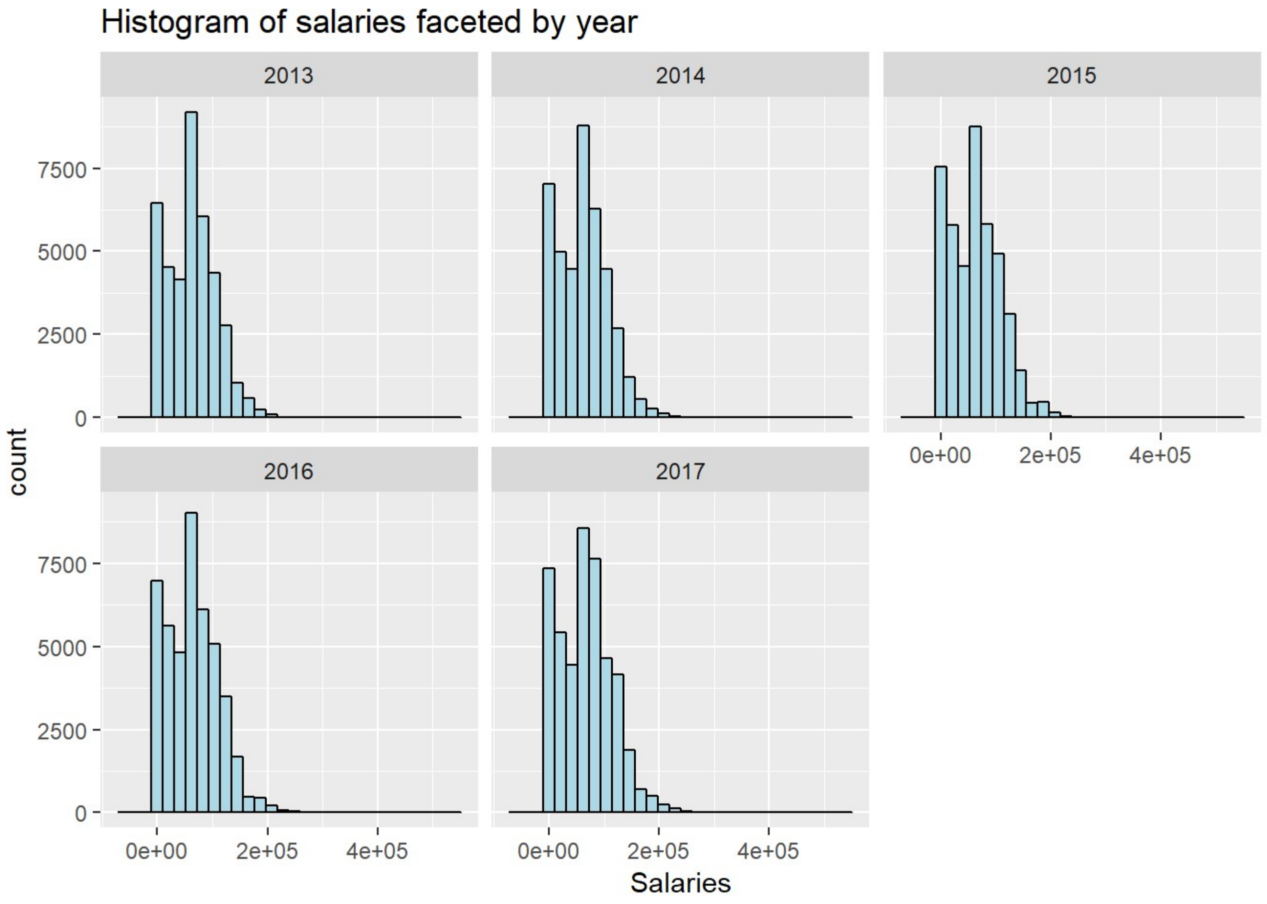
```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot()+geom_boxplot(mapping = aes(x=Year,y=Salaries, group=Year), data=Emp_Comp)+scale_y_continuous(limits=c(0,150000),breaks=seq(0,150000,25000))+ggtitle("Box plot of salaries per year - closer look")
```



1b) Histograms faceted by year.

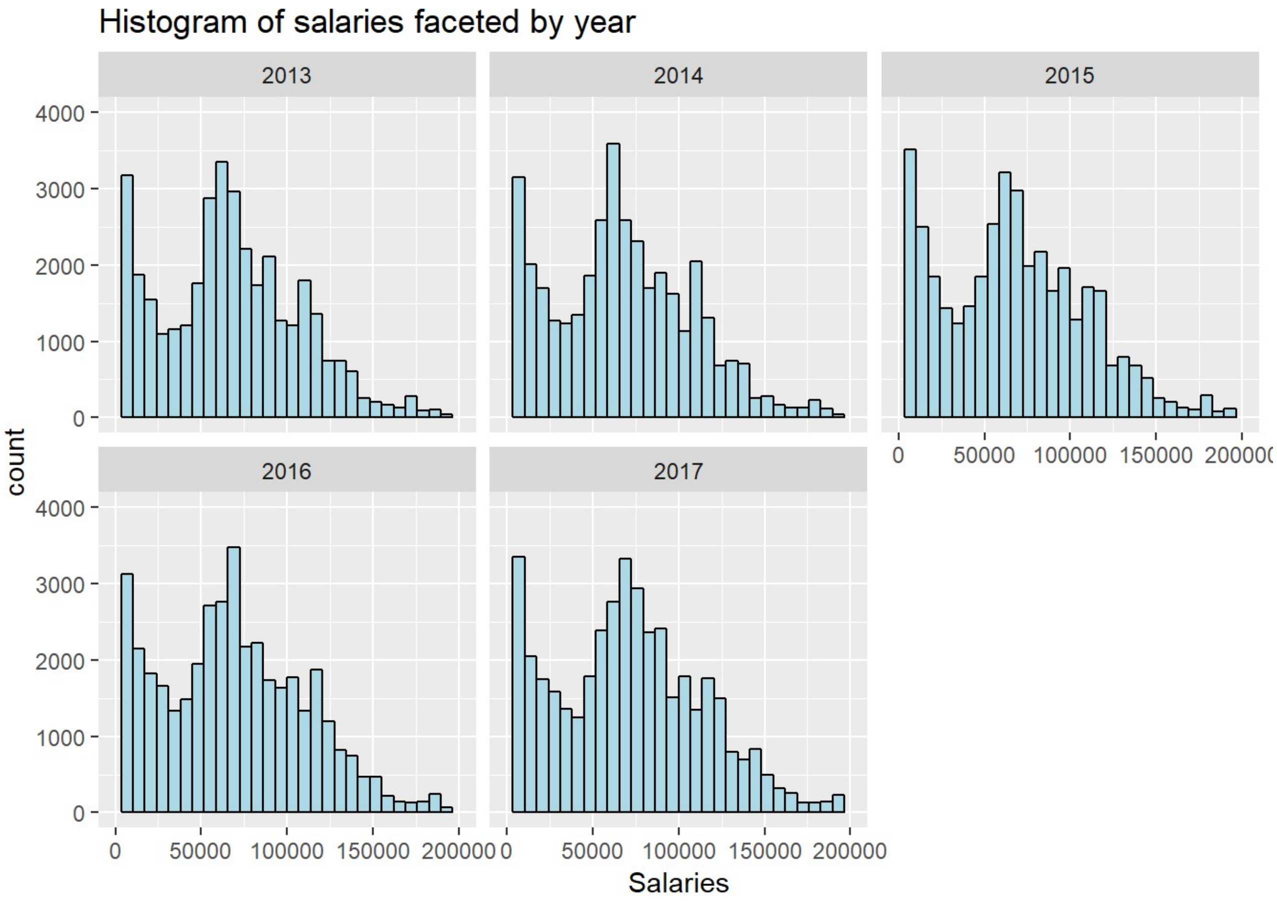
A histogram is useful to study the density of data points in a data set. The salaries are shown on the x axis and number of occurences of a particular salary range is plotted on the y axis. The width of each bin in the histogram denotes the interval of data contained in that bin.

```
ggplot(Emp_Comp,aes(x=Salaries))+geom_histogram(color="black",fill="lightblue")+facet_wrap(. ~ Year)+ggtitle("Histogram of salaries faceted by year")
```



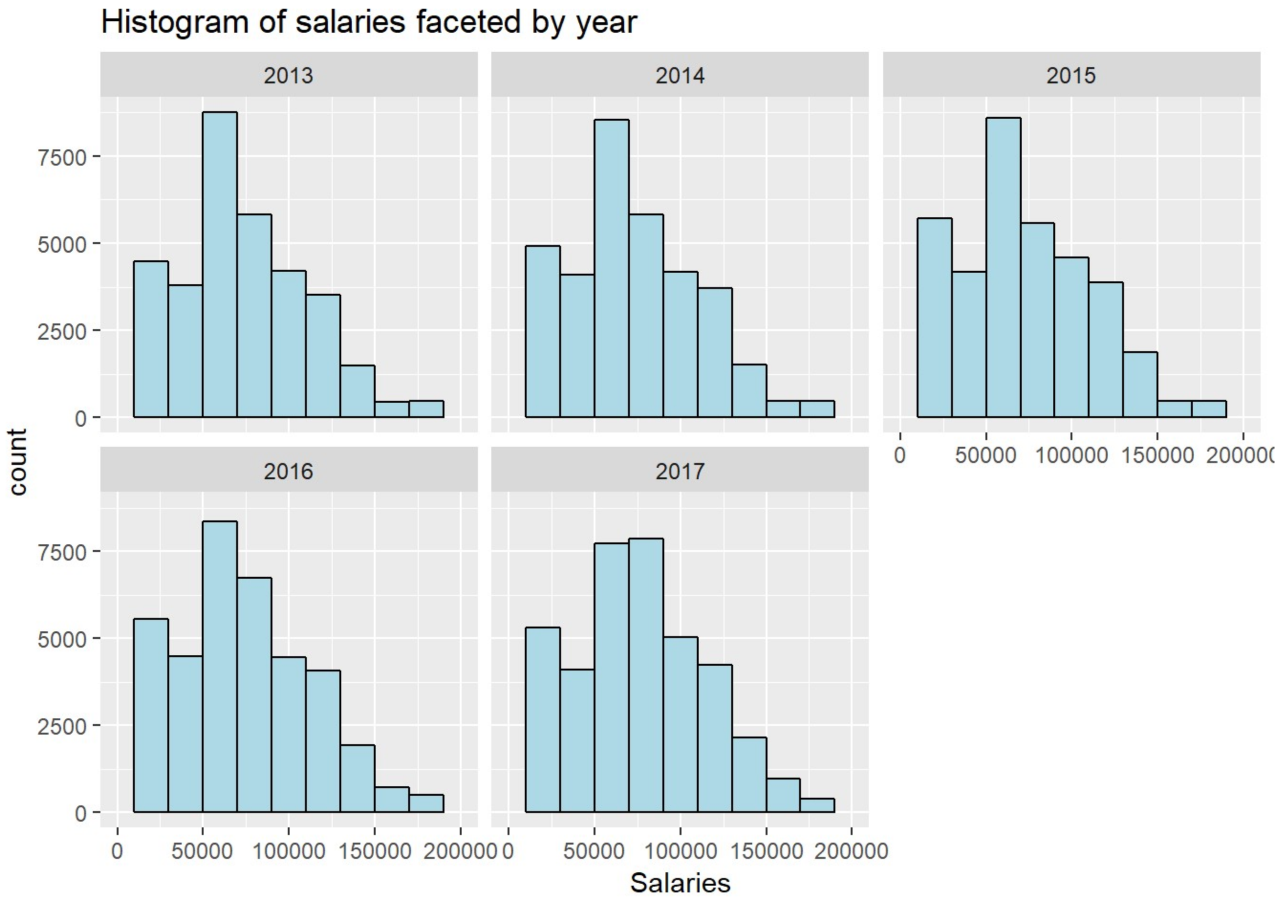
We see that histograms are bad at showing the outliers. Hence we will limit the x axis to around 200000 so that we get better view of the data.

```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot(Emp_Comp,aes(x=Salaries))+geom_histogram(color="black",fill="lightblue")+facet_
wrap(. ~ Year) + xlim(0, 200000)+ggtitle("Histogram of salaries faceted by year")
```



From the dataset, we see that salaries are mentioned for each “Job”. So, y axis in histogram of Salaries can be interpreted as the number of jobs that fall in that particular salary range. By changing the binwidth, we can view the plot with various granularities.For example,

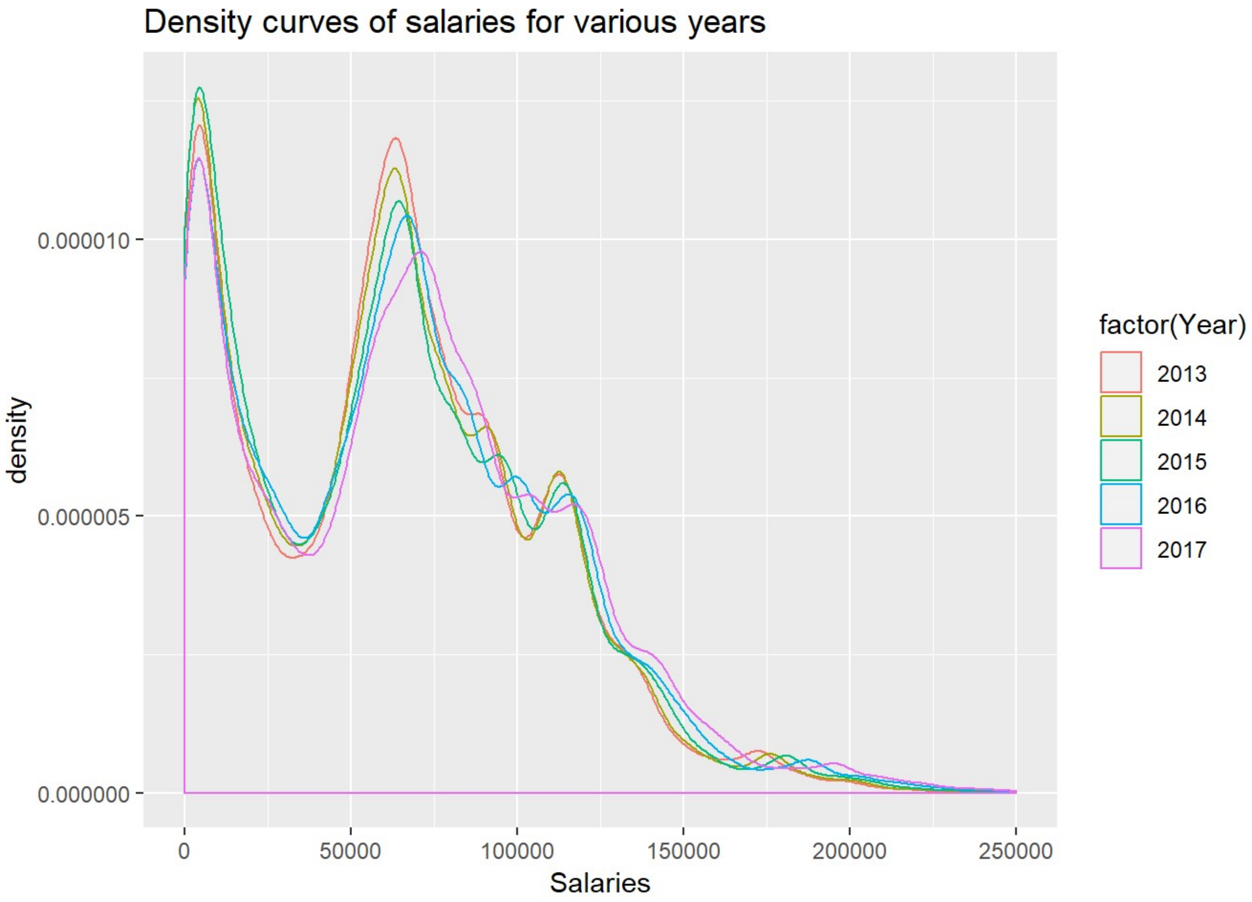
```
Emp_Comp <- read.csv("Employee_Compensation.csv")
options(scipen=999)
ggplot(Emp_Comp,aes(x=Salaries))+geom_histogram(color="black",fill="lightblue",binwidth=20000) + facet_wrap(. ~ Year) + scale_x_continuous(limits = c(0,200000),breaks = seq(0,200000,50000))+ggtitle("Histogram of salaries faceted by year")
```



1c) Plotting density curves Density curves are useful in understanding the distribution of data in an interval.The peaks represent high density regions.i.e.more number of jobs are salaried at the particular value. Density plot represents the model of the data and hence can be used in probability applications.Multi modality is also seen in density curves.

The following plot can be used to compare the salary denisities from year to year. Since they are overlaid on one another on the same graph, comparison is easier than faceted histograms.

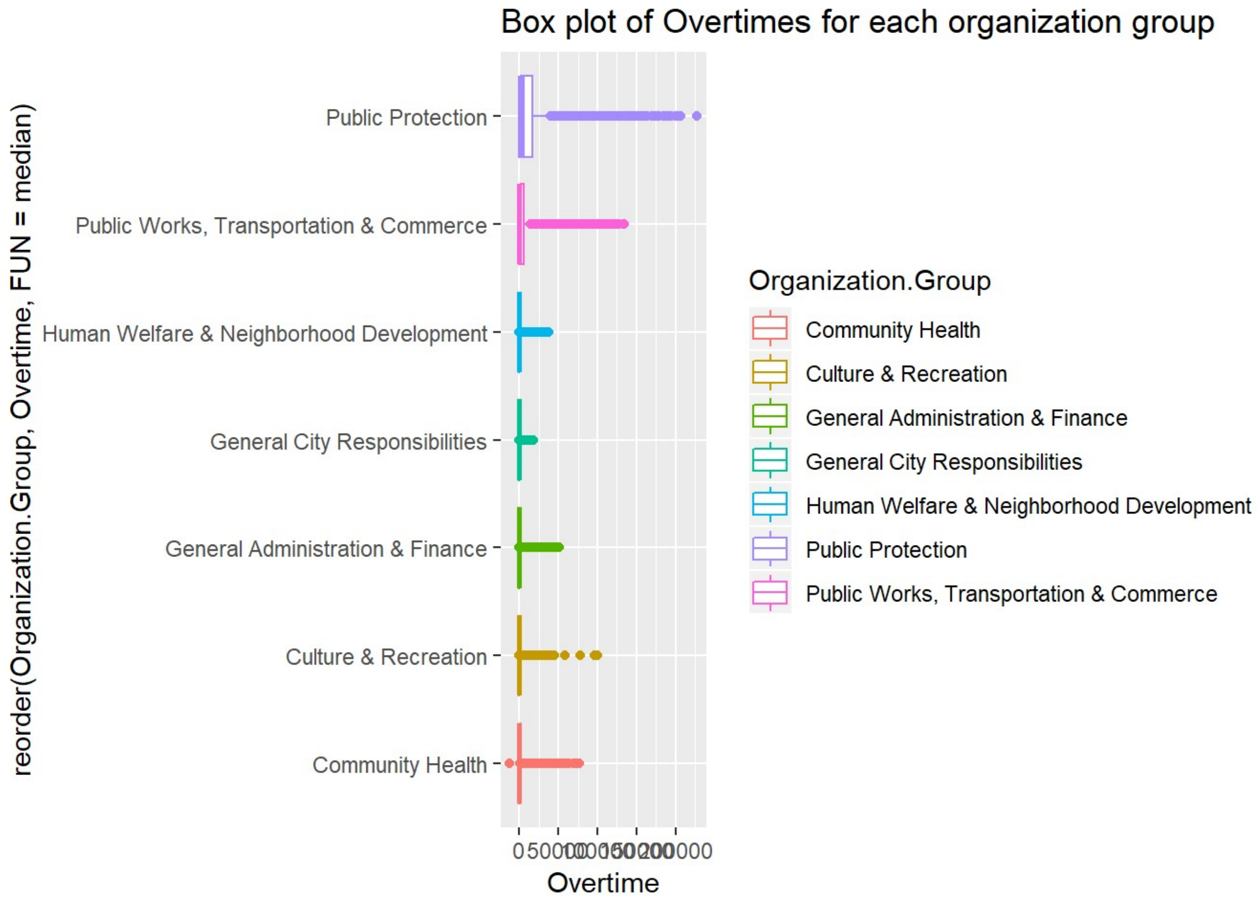
```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot(Emp_Comp,aes(x=Salaries,color=factor(Year)))+geom_density()+xlim(0,250000)+ggti
tle("Density curves of salaries for various years")
```



1d) The questions that can be answered from the above plots are: From boxplot: 1) What is the median salary for jobs in each year? 2) Find the salary outliers in each year From histogram: 1) How many jobs get salaries in the range(x,y)? 2) What is the salary range that most job categories are paid? 3) Compare the number of jobs that are paid same salary range from all the years. 4) How is the spread of the data? Are there any unusual peaks? From density curve: 1) How do you model the data on seeing the plot? 2) Draw comparision of peaks since it contains overlapping curves for all the years in one plot. 3) Determine the probabilistic model for the data in each year.

2a) Horizontal boxplot in terms of sorted median. The plot shows that a lot of data points are outliers and hence this plot is not good for representing the data.

```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot(Emp_Comp, aes(x=reorder(Organization.Group,Overtime,FUN=median),y=Overtime, col
or=Organization.Group)) + geom_boxplot()+coord_flip()+ggtitle("Box plot of Overtimes f
or each organization group")
```



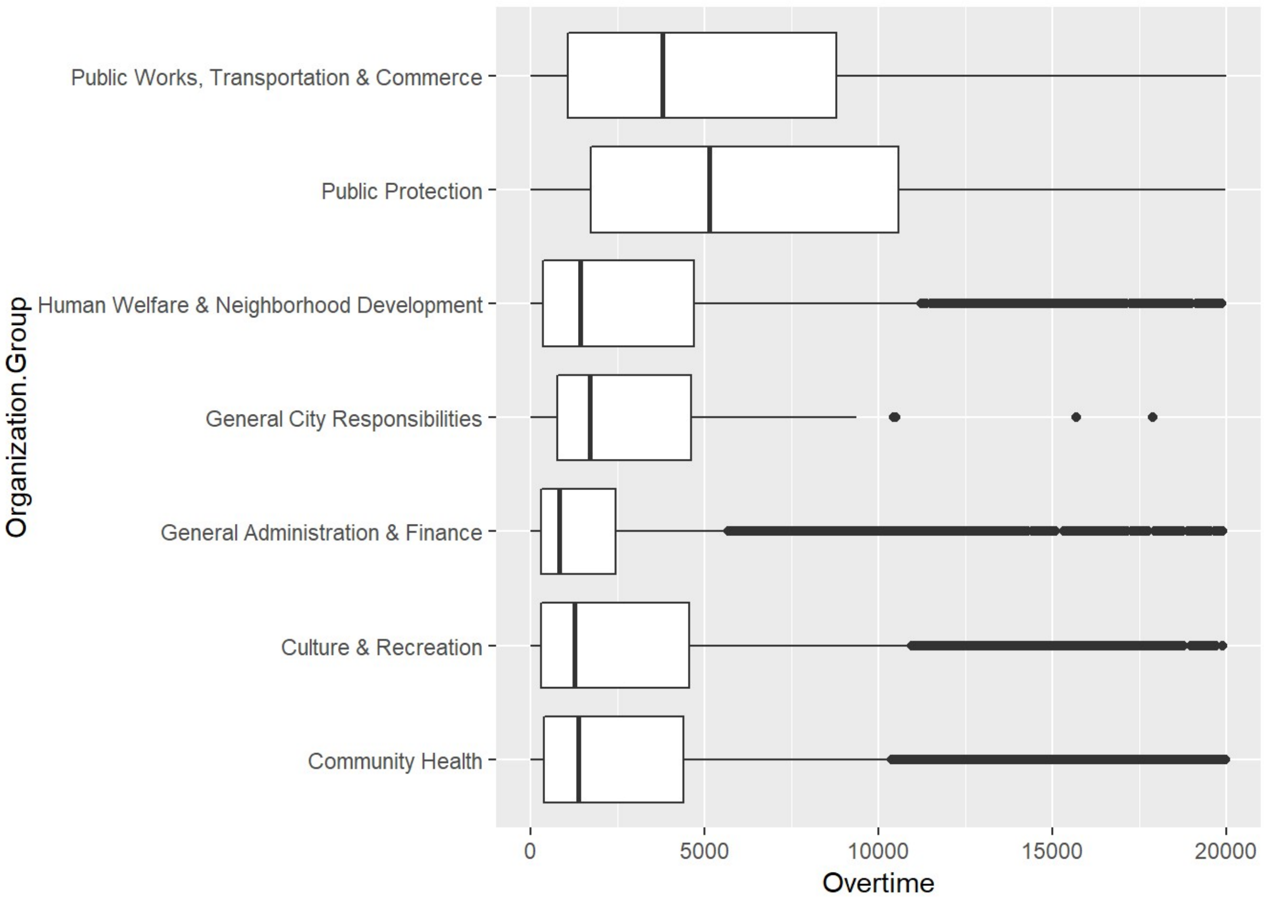
2b) Subsetting the data to get better visualization. Dplyr comes in handy when we need to subset data based on the range of values in a particular column. Lets subset the data into 3 parts based on overtime ranges as shown below. On subsetting, we get a better view of the distribution.

Note: Negative overtime values are ignored while subsetting the data

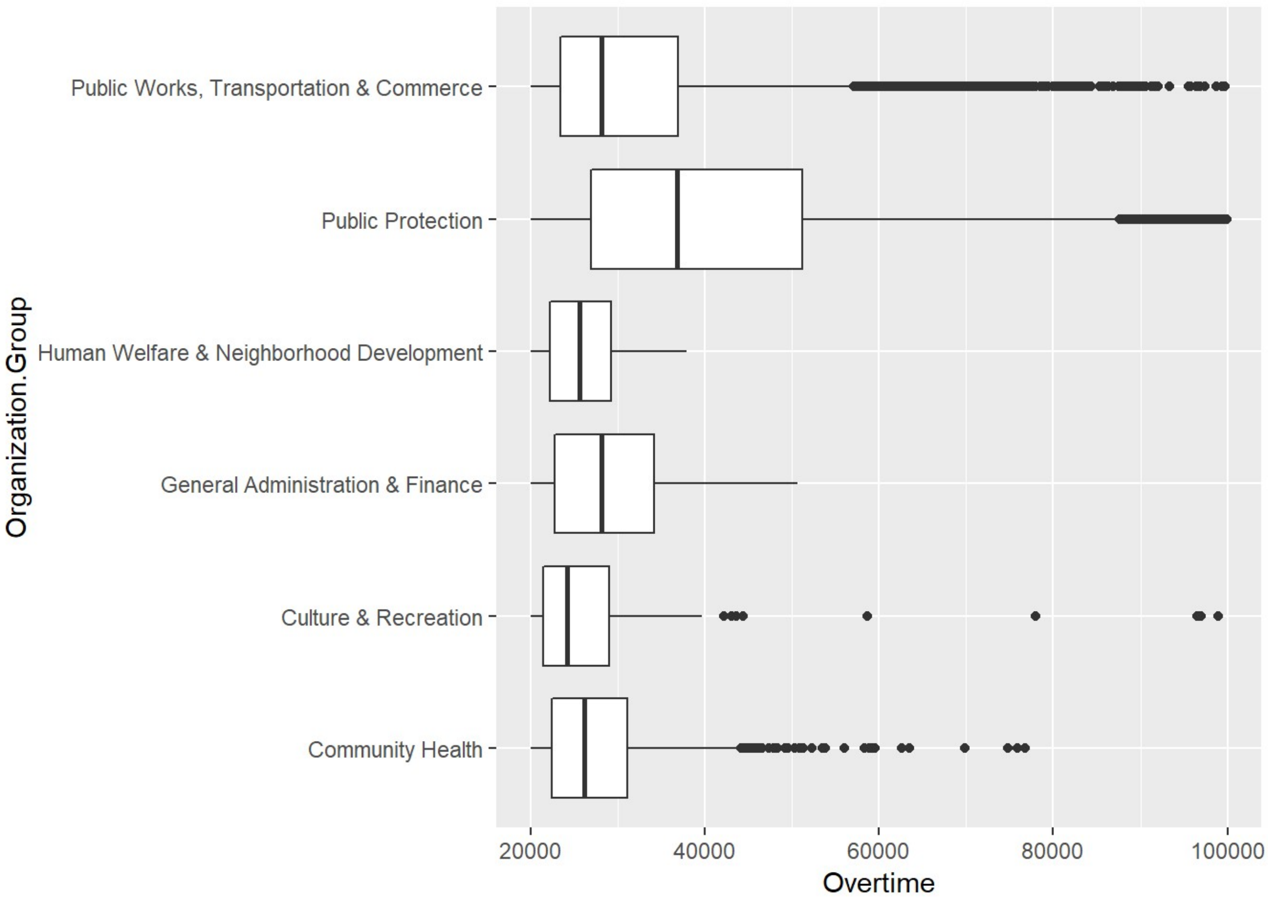
We get to understand the quartiles, range, median and outliers for each of the subsets considered as separate sets.

In the third plot, there are only 3 organizations because rest of the organizations do not have overtime values lying in that range.

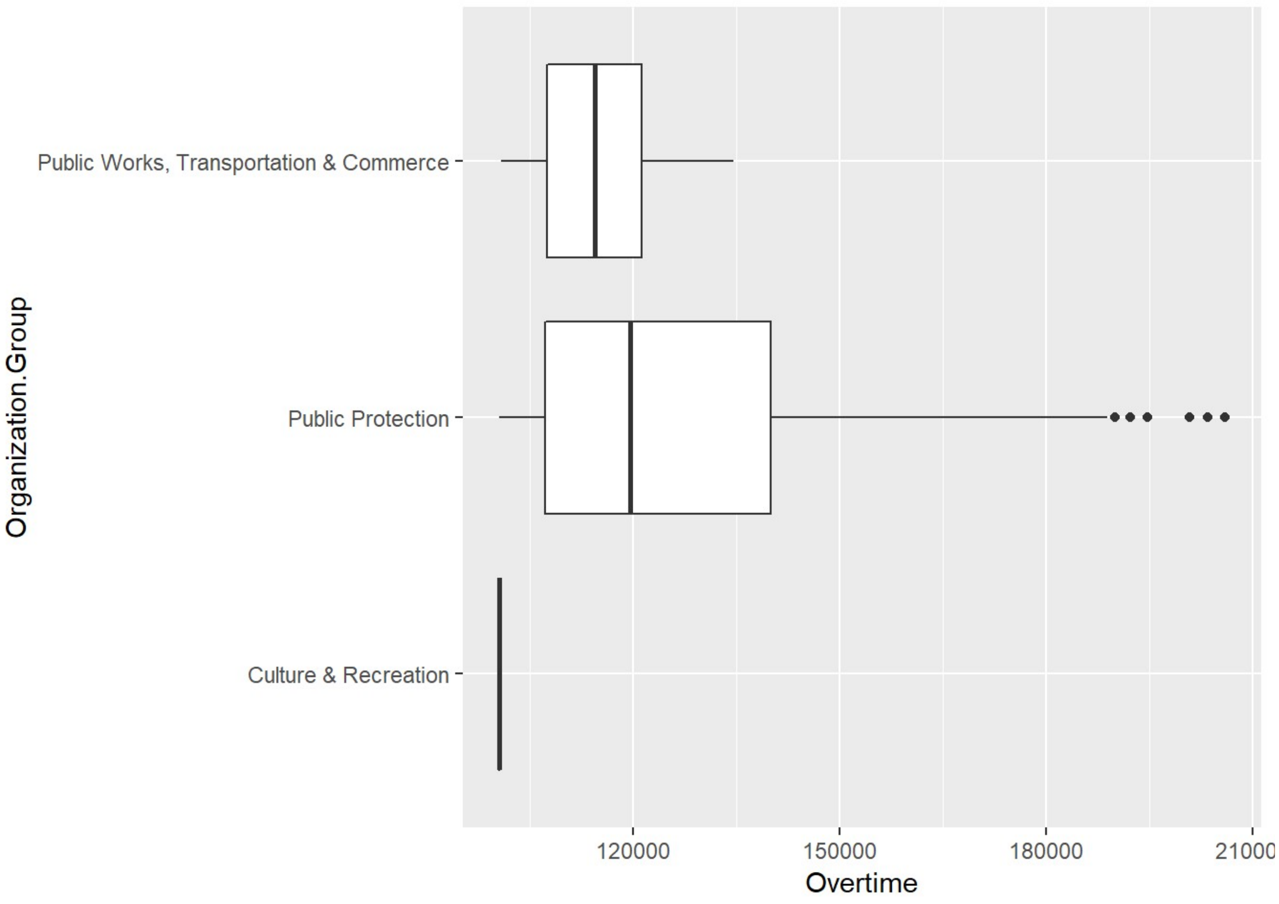
```
Emp_Comp <- read.csv("Employee_Compensation.csv")
options(scipen=999)

subset_data <- Emp_Comp %>% filter(Overtime>0 & Overtime<20000) %>% select(Overtime,Organization.Group)
ggplot(subset_data, aes(x=Organization.Group,y=Overtime))+geom_boxplot()+coord_flip()
```



```
subset_data1 <- Emp_Comp %>% filter(Overtime>20000 & Overtime<100000) %>% select(Overt
ime,Organization.Group)
ggplot(subset_data1, aes(x=Organization.Group,y=Overtime))+geom_boxplot()+coord_flip()
```



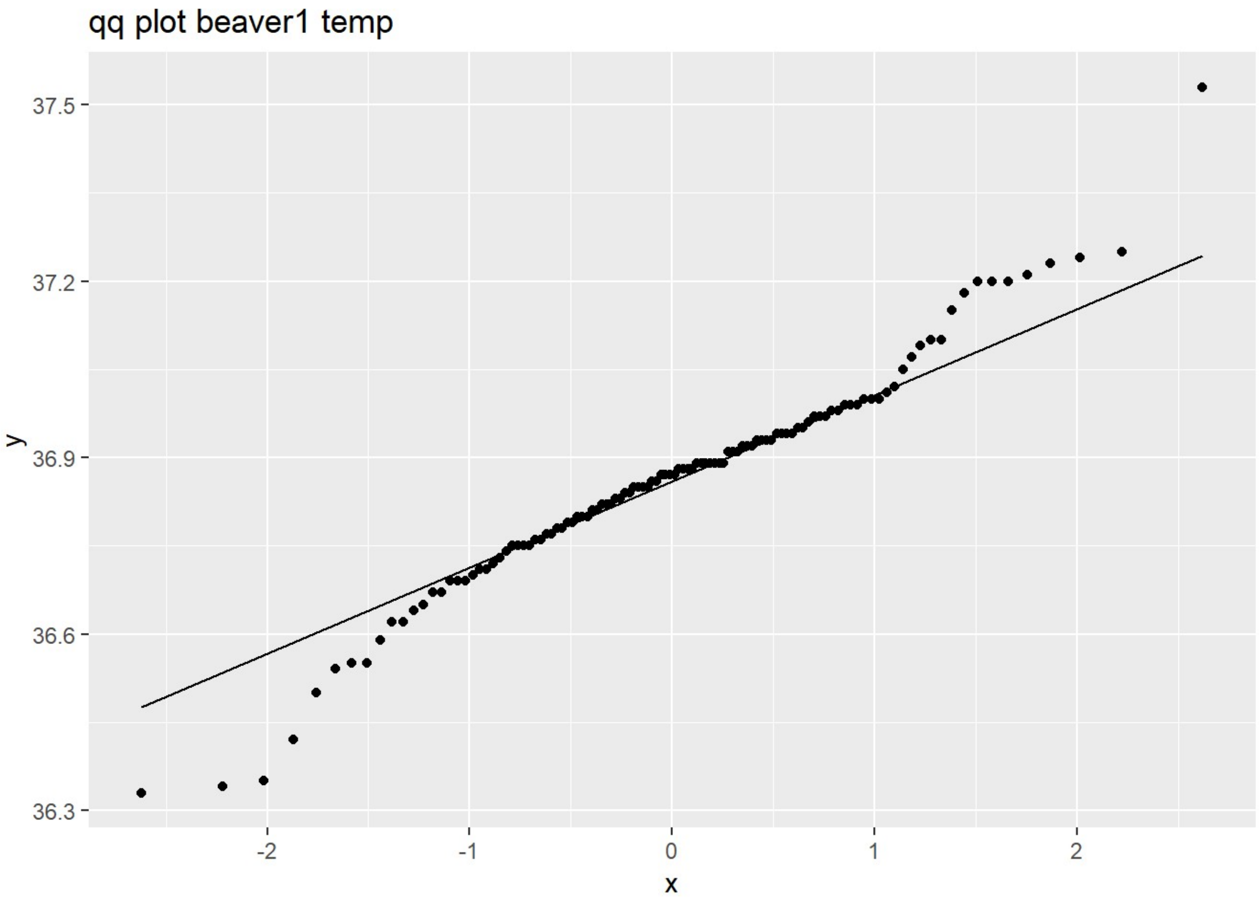
```
subset_data2 <- Emp_Comp %>% filter(Overtime>100000 & Overtime<227313) %>% select(Over
time,Organization.Group)
ggplot(subset_data2, aes(x=Organization.Group,y=Overtime))+geom_boxplot()+coord_flip()
```



4a) A QQ plot is used for two purposes: to determine visually whether a given dataset is normal or to compare the distribution of two datasets. It plots the dataset on y axis and theoretical normal dataset on x axis. If the points fall in a line such that x=y, then it means that input dataset matches with theoretical normal dataset and hence the input is normal.

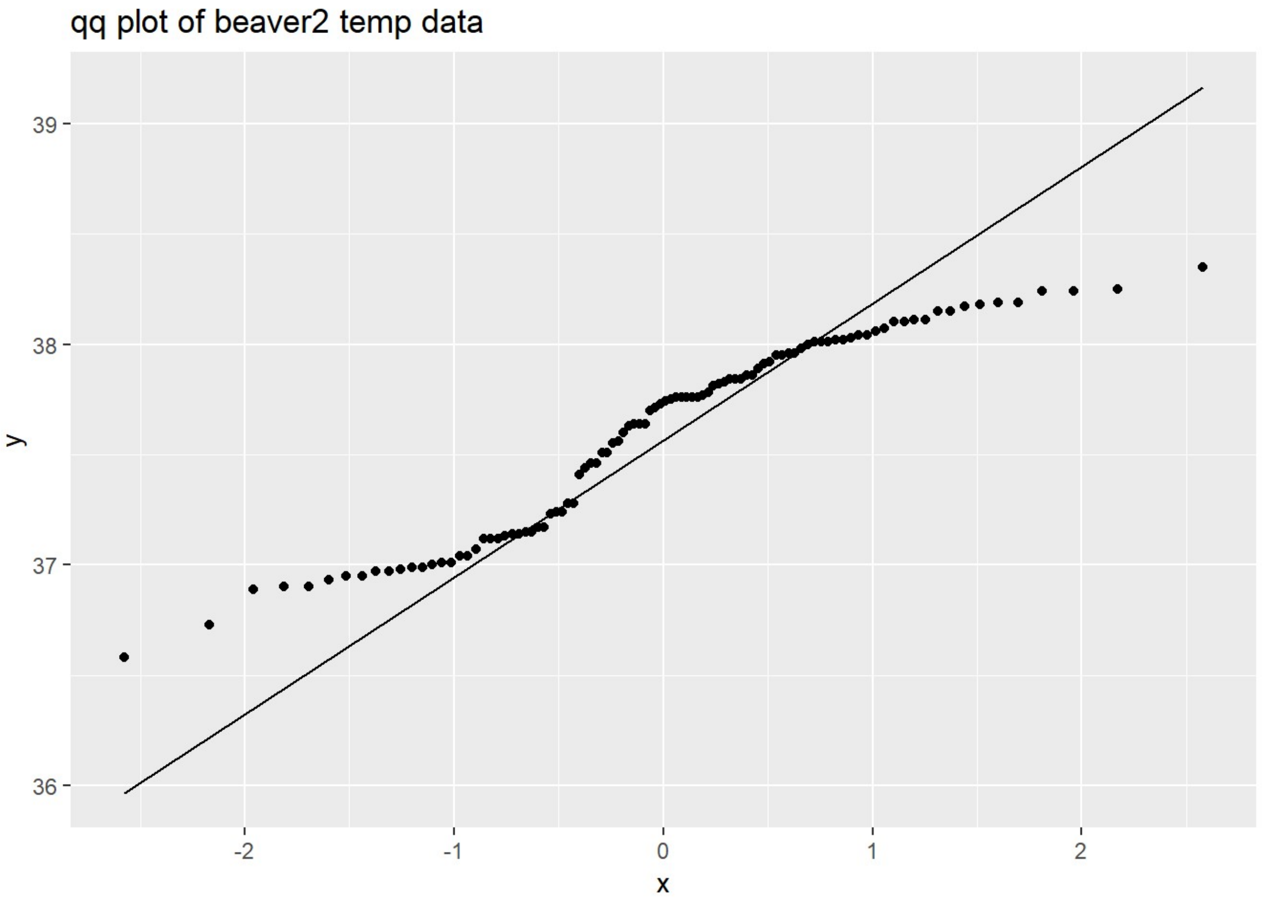
QQ plot for beaver1 suggests that the temp is approximately normally distributed.

```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot(beaver1,aes(sample=temp))+geom_qq_line()+geom_qq()+ggtitle("qq plot beaver1 tem
p")
```



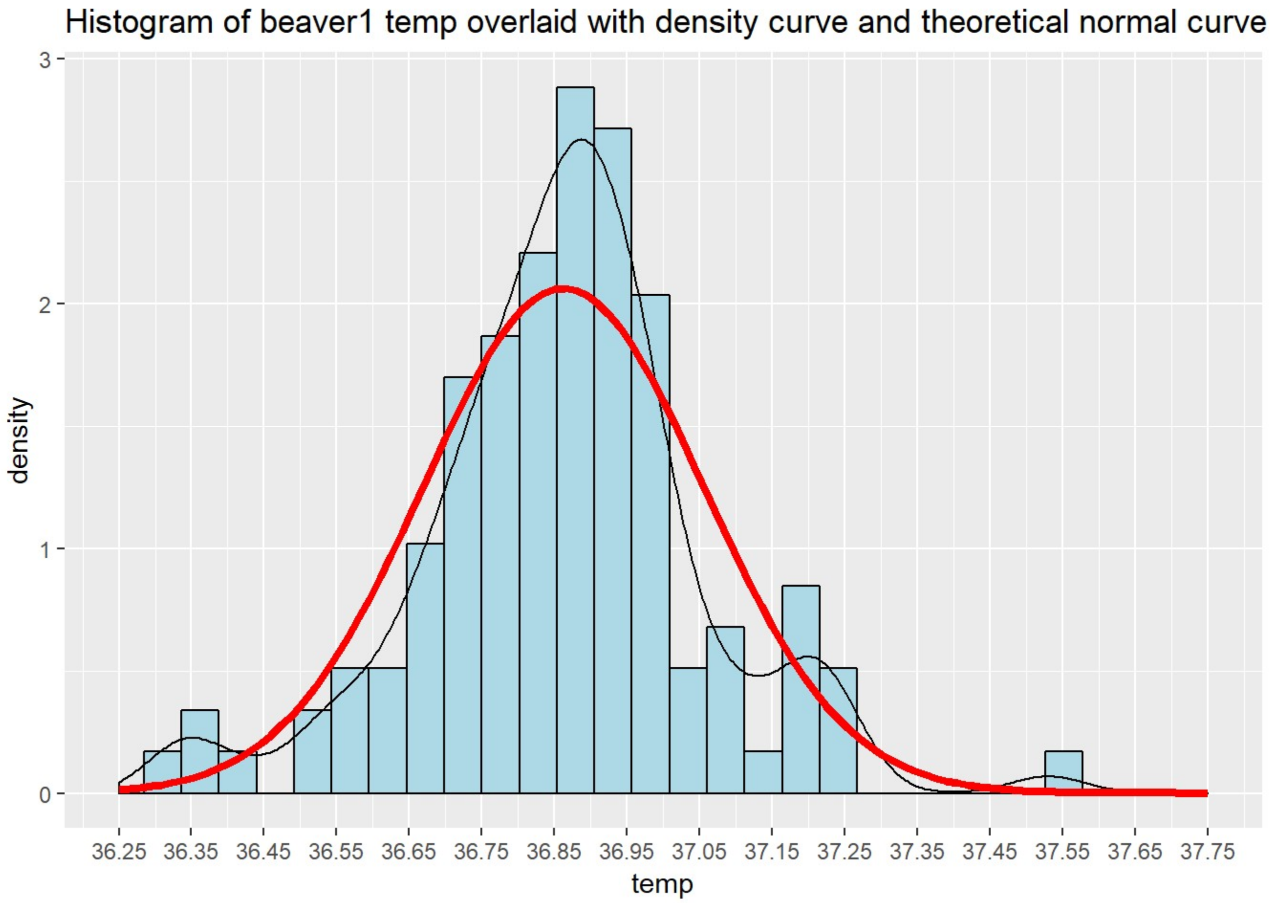
QQ plot of beaver2 suggests that the temp is not normally distributed. It may be bi-modal

```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot(beaver2,aes(sample=temp))+geom_qq_line()+geom_qq()+ggtitle("qq plot of beaver2
temp data")
```



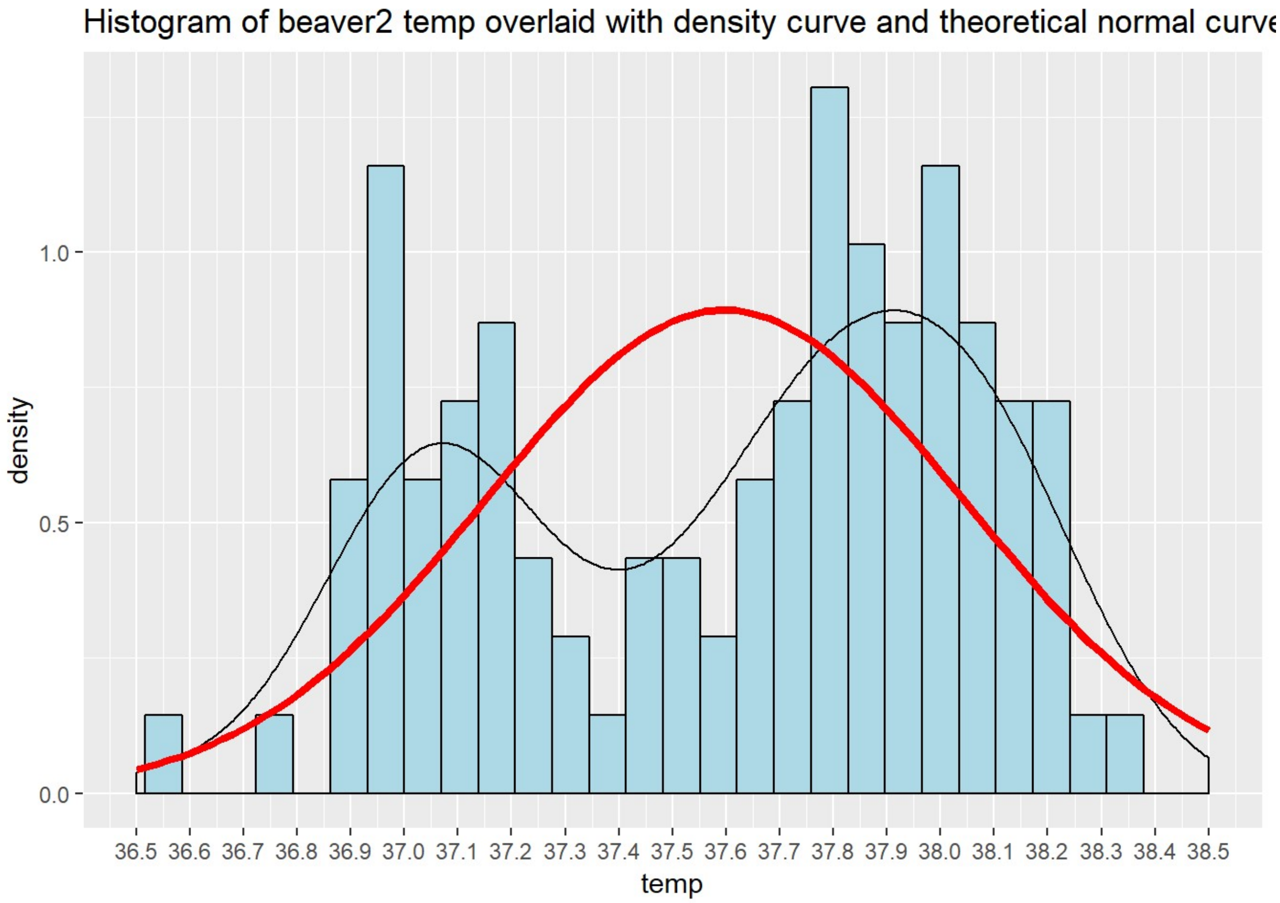
4b) Beaver1 histogram with curve overlaid shows that the distribution is approximately normal with a small increase in density at temperature 37.20

```
ggplot(beaver1, aes(x=temp))+geom_histogram(aes(y=..density..),color="black",fill="lightblue")+geom_density()+scale_x_continuous(limits=c(36.25,37.75),breaks = seq(36.25,37.75,0.1))+stat_function(fun=dnorm, args=list(mean=mean(beaver1$temp),sd=sd(beaver1$temp)),color="red",lwd=1.5)+ggtitle("Histogram of beaver1 temp overlaid with density curve and theoretical normal curve")
```



Beaver2 is not normal as seen from the density and histogram plot below.

```
Emp_Comp <- read.csv("Employee_Compensation.csv")
ggplot(beaver2, aes(x=temp))+geom_histogram(aes(y=..density..),color="black",fill="lightblue")+geom_density()+scale_x_continuous(limits=c(36.5,38.5),breaks = seq(36.5,38.5,0.1))+stat_function(fun=dnorm, args=list(mean=mean(beaver2$temp),sd=sd(beaver2$temp)),color="red",lwd=1.5)+ggtitle("Histogram of beaver2 temp overlaid with density curve and theoretical normal curve")
```



4c) Shapiro-wilk test on beaver1 and beaver2 show that p-value is extremely small for beaver2 dataset and hence the null hypothesis (data is normal) is rejected for beaver2. For beaver1, the null hypothesis holds.

```
shapiro.test(beaver1$temp)
```

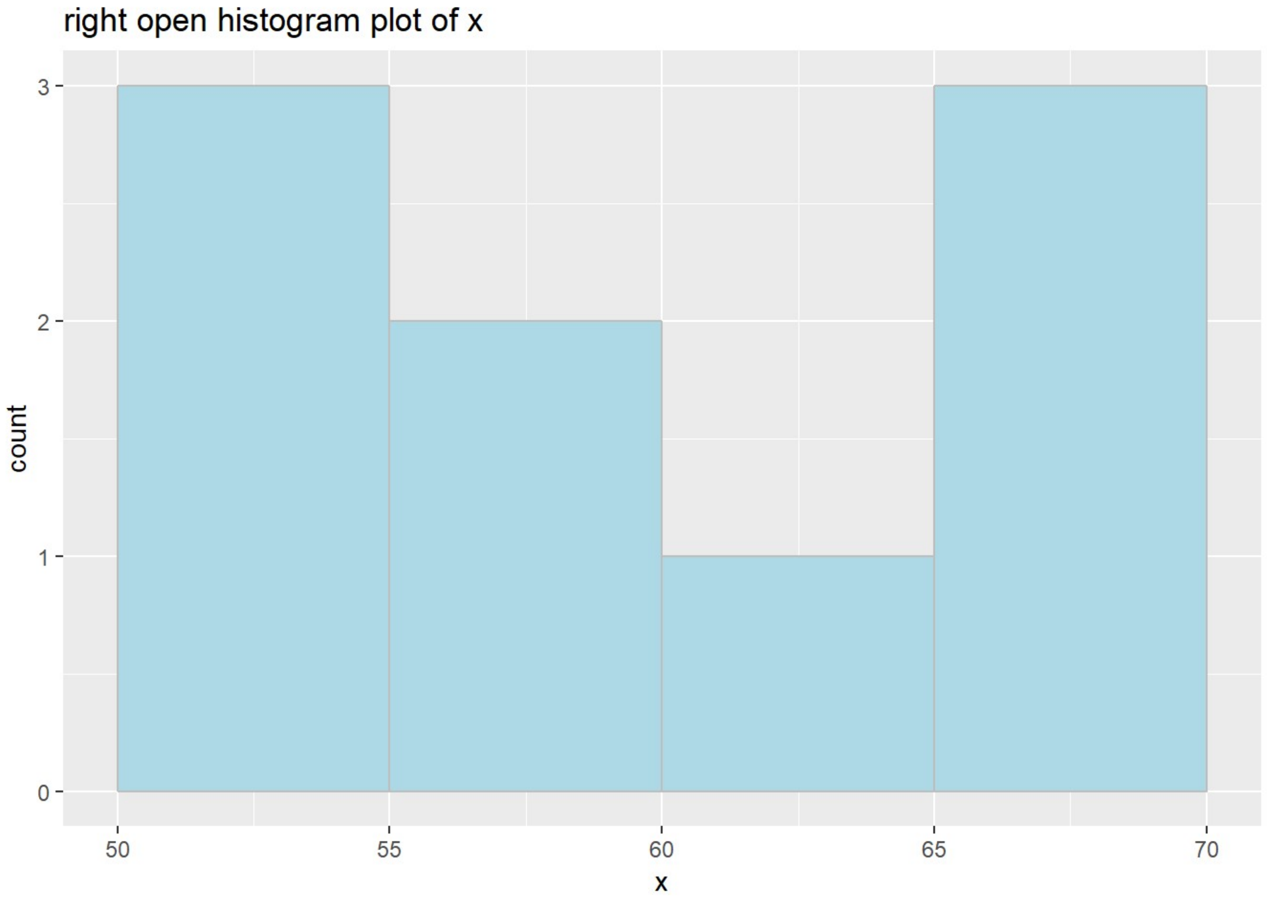
```
##
##  Shapiro-Wilk normality test
##
## data:  beaver1$temp
## W = 0.97031, p-value = 0.01226
```

```
shapiro.test(beaver2$temp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beaver2$temp
## W = 0.93336, p-value = 0.00007764
```

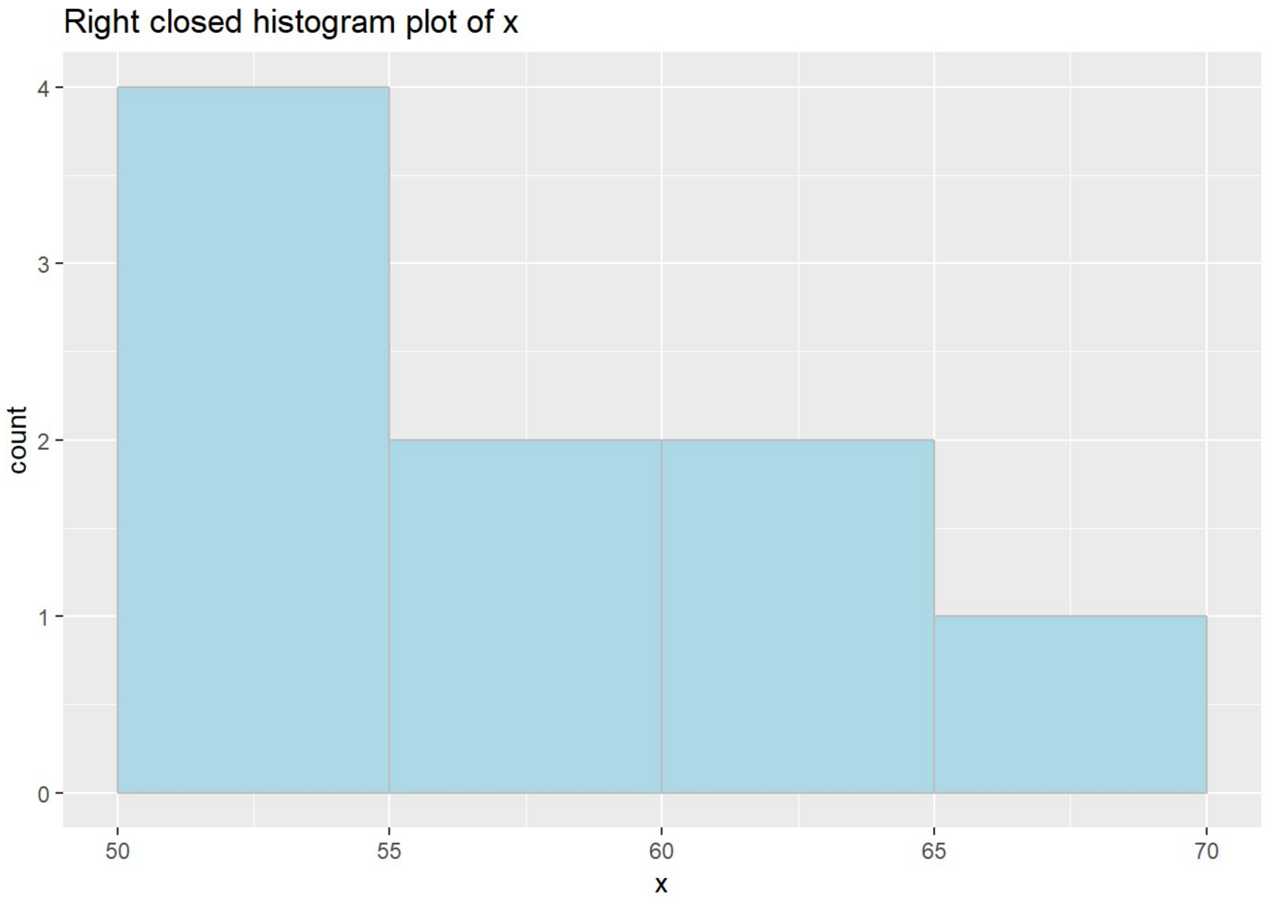
3a)Right open histogram

```
x <- c(50, 51, 53, 55, 56, 60, 65, 65, 68)
df = data.frame(x)
ggplot(df, aes(x))+geom_histogram(color = "grey", fill= "lightBlue",binwidth=5,center=
52.5,closed=c("left"))+ggtitle("right open histogram plot of x")
```



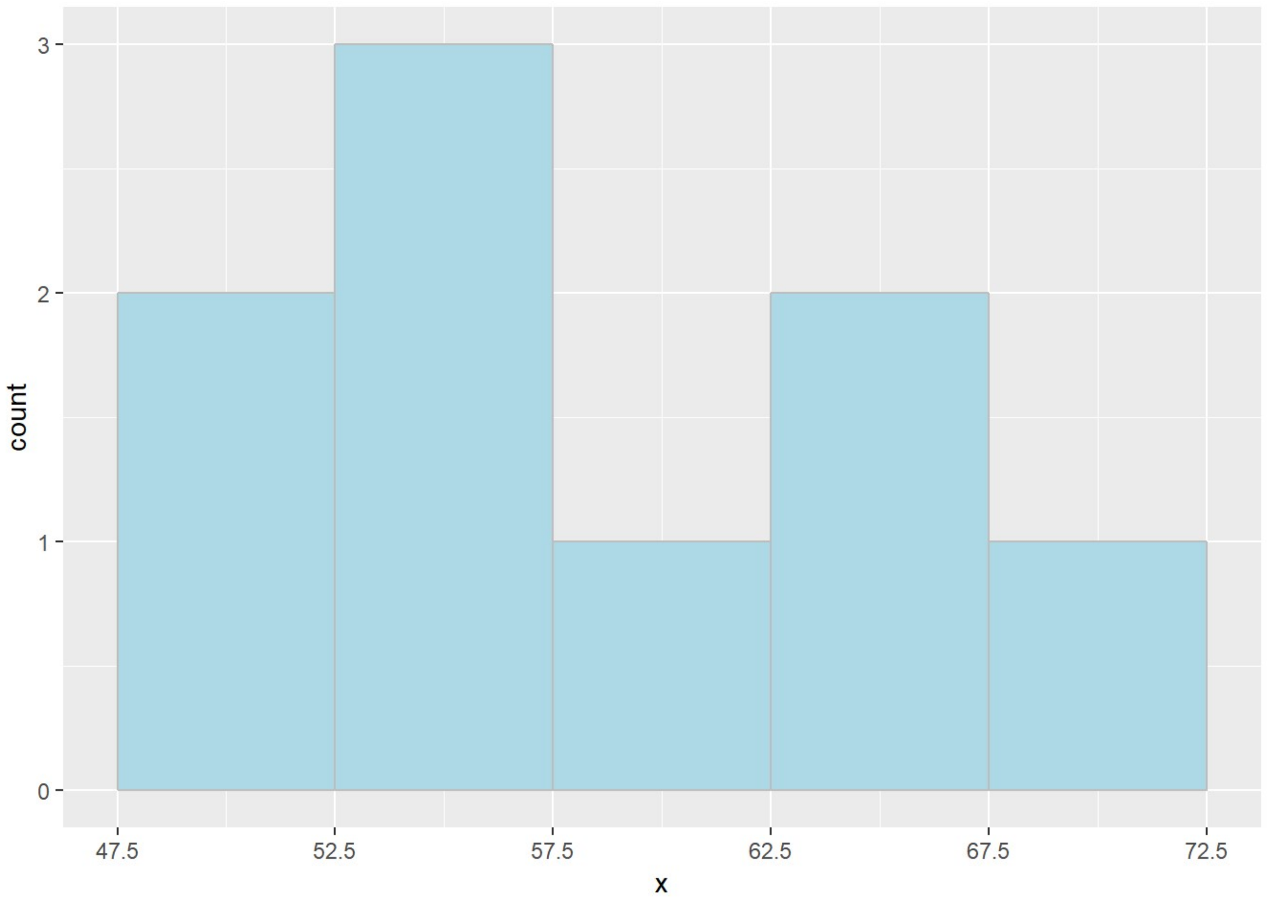
Right closed histogram

```
x <- c(50, 51, 53, 55, 56, 60, 65, 65, 68)
df = data.frame(x)
ggplot(df, aes(x))+geom_histogram(color = "grey", fill = "lightBlue",binwidth=5,center
=52.5,closed=c("right"))+ggtitle("Right closed histogram plot of x")
```

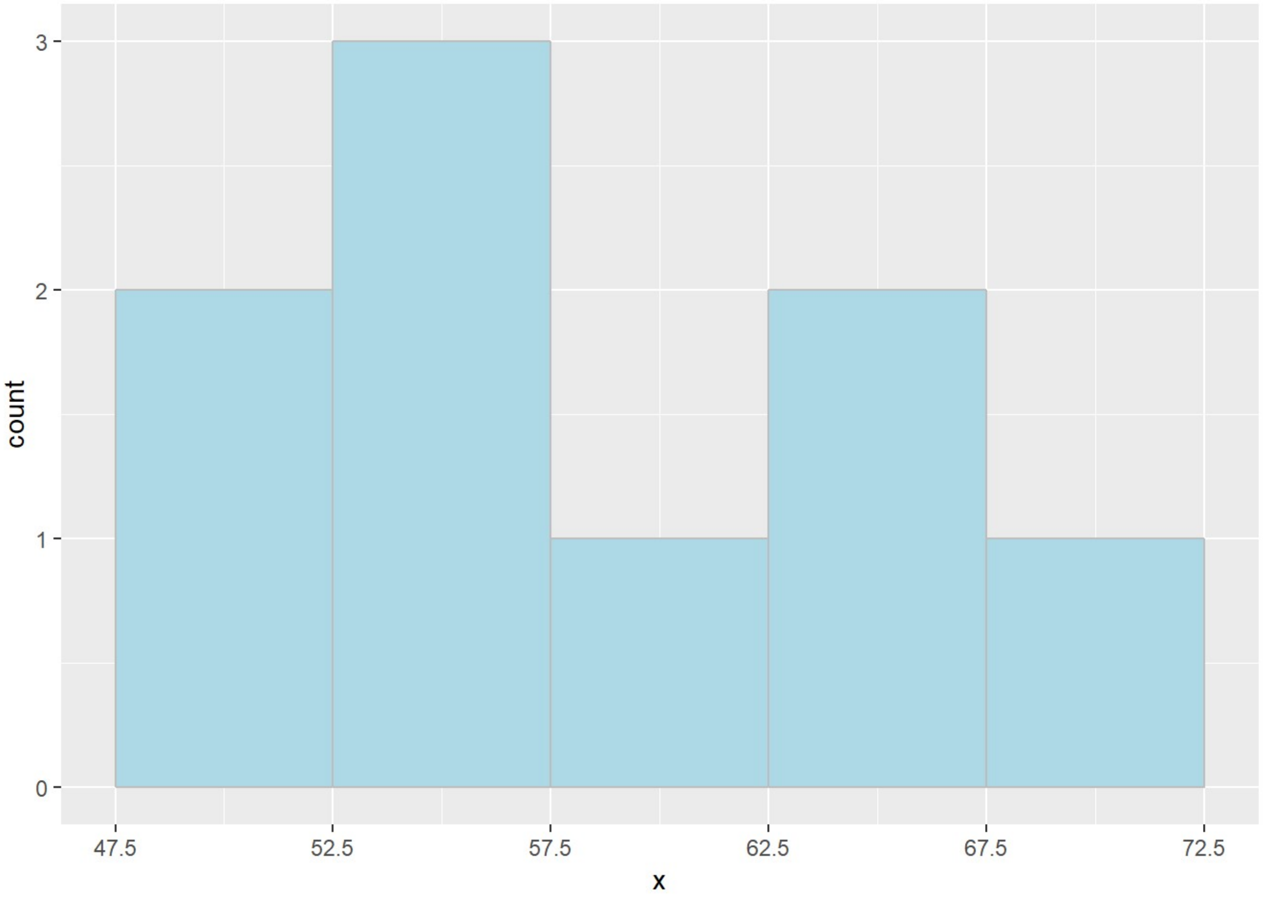


Making both identical by changing the bin boundaries to 47.5-72.5. In this manner, the data points don't fall in the boundary and hence left or right open interval doesn't matter

```
x <- c(50, 51, 53, 55, 56, 60, 65, 65, 68)
df = data.frame(x)
ggplot(df, aes(x))+geom_histogram(color = "grey", fill = "lightBlue",binwidth=5,closed
=c("left"))+scale_x_continuous(limits=c(47.5,72.5),breaks=seq(47.5,72.5,5))
```



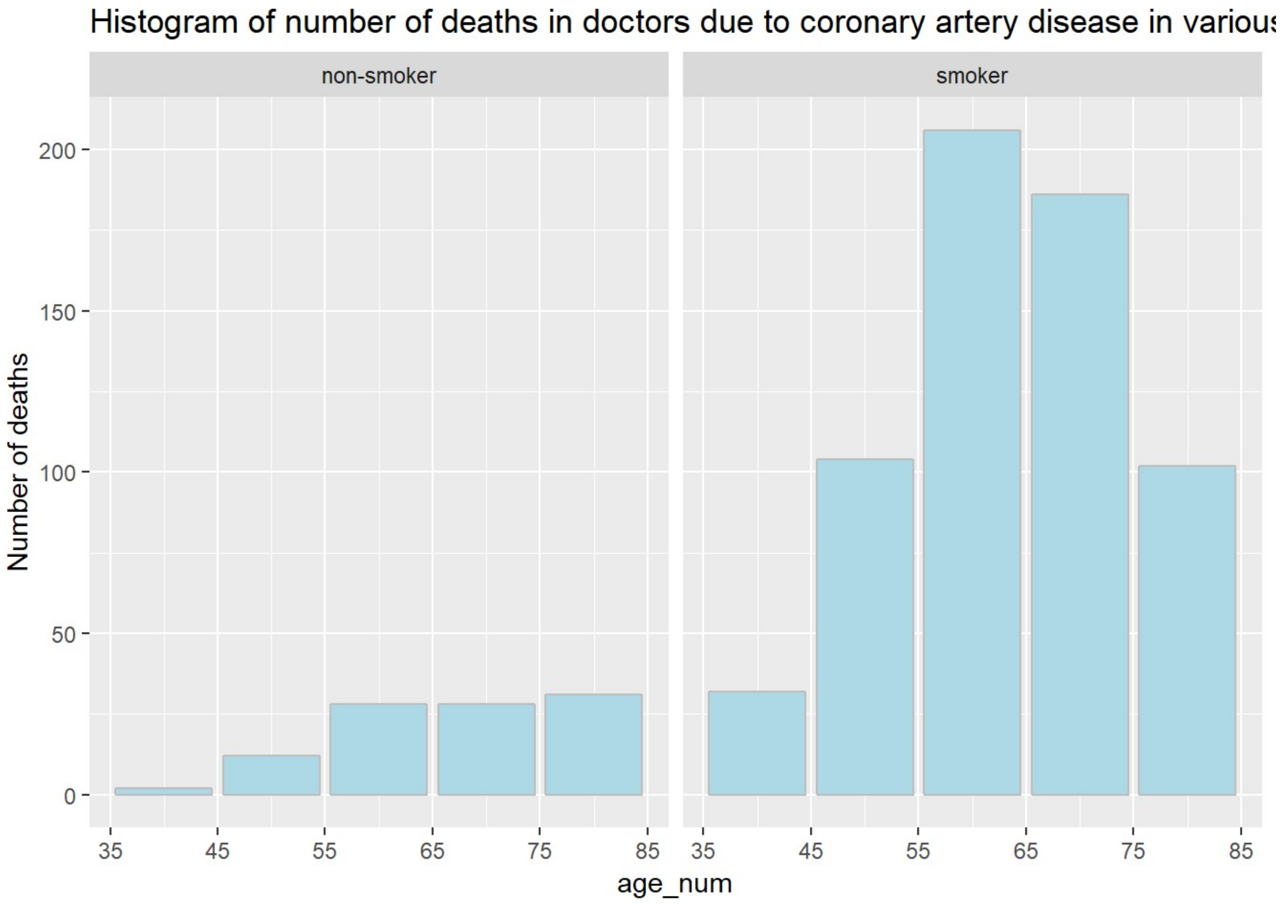
```
ggplot(df, aes(x))+geom_histogram(color = "grey", fill = "lightBlue",binwidth=5,closed
=c("right"))+scale_x_continuous(limits=c(47.5,72.5),breaks=seq(47.5,72.5,5))
```



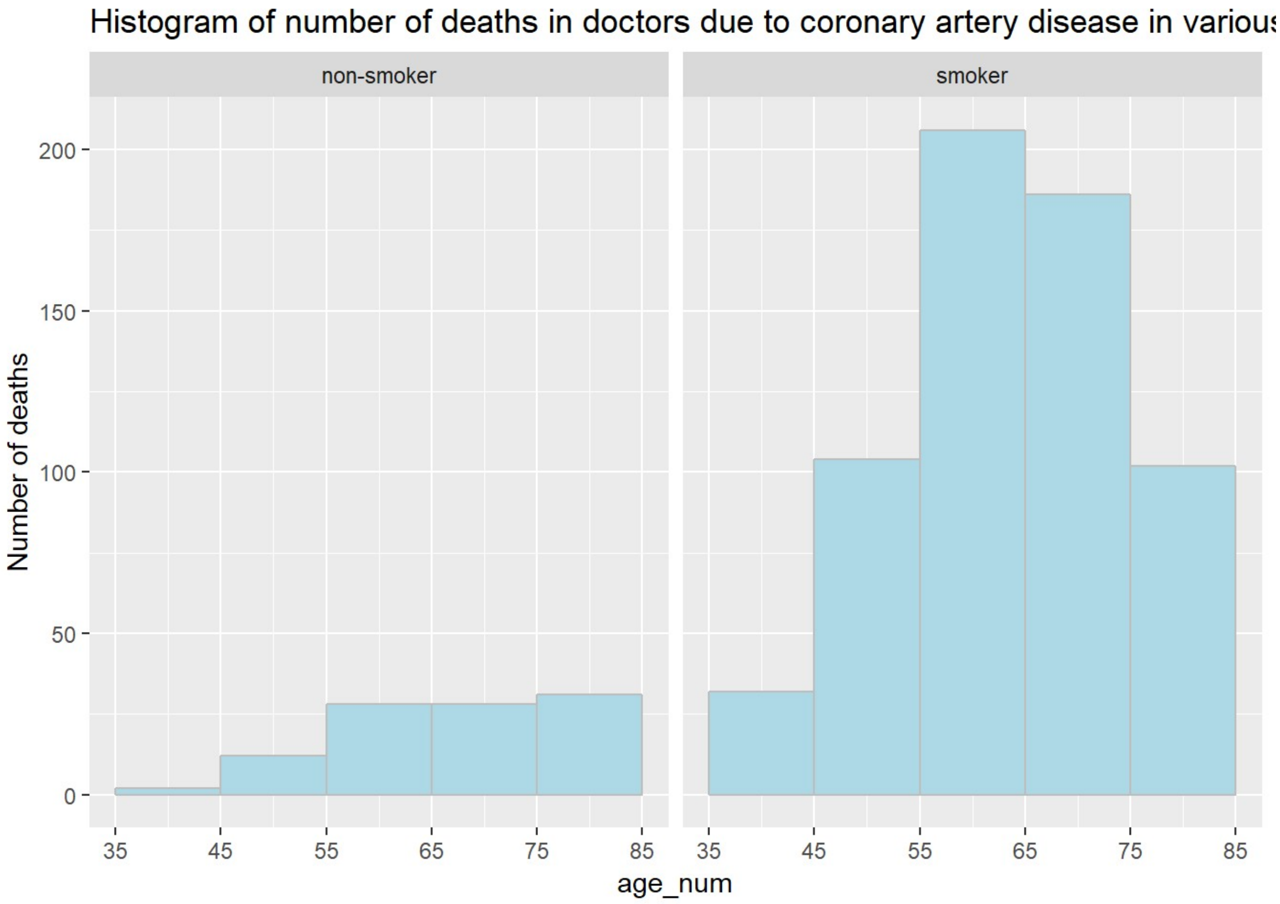
5. The following is the histogram of the number of deaths caused by coronary artery disease among doctors faceted by whether they are smokers or non-smokers.

The histogram clearly shows gaps in between the bars on the x axis. This is because age is not a continuous variable. It is discrete in our dataset. According to breslow dataset, the age is a factor and is the midpoint of 10 year range.

```
new_breslow <- breslow #copy into new variable as the original dataset remains undistu
rbed

new_breslow$smoke <- factor(new_breslow$smoke)


new_breslow$age_num = as.numeric(levels(new_breslow$age))[new_breslow$age]
#new_breslow

levels(new_breslow$smoke) <- c("non-smoker","smoker")
ggplot(new_breslow,aes(x=age_num,y=y))+geom_histogram(color = "grey",stat="identity",f
ill = "lightBlue")+facet_wrap(~smoke)+labs(y="Number of deaths") + scale_x_continuous
(breaks=c(35,45,55,65,75,85)) + ggtitle("Histogram of number of deaths in doctors due
to coronary artery disease in various age ranges")
```



```
#Using geom_col
ggplot(new_breslow,aes(x=age_num,y=y))+geom_col(width=10,color = "grey",fill = "lightB
lue",stat="identity")+facet_wrap(~smoke)+labs(y="Number of deaths", binwidth=50)+scale
_x_continuous(breaks=c(35,45,55,65,75,85)) + ggtitle("Histogram of number of deaths i
n doctors due to coronary artery disease in various age ranges")
```



Histogram of number of deaths in doctors due to coronary artery disease in various